統計II（医療統計II）
第2回　母平均と母比率の区間推定

授業担当：徳永伸一

東京医科歯科大学教養部　数学講座

あらためて：第13章 推定　の内容

Ⅰ．母集団と標本
Ⅱ．点推定
- 不偏性，不偏推定量

Ⅲ．区間推定
Ⅳ．母平均の区間推定
- 母分散が既知のとき

＊＊＊前期試験範囲ここまで＊＊＊

- 母分散が未知のとき

Ⅴ．母分散の区間推定
Ⅵ．母比率の区間推定

第14章 検定　の内容

Ⅰ．検定
Ⅱ．母平均の検定
Ⅲ．母分散の検定
Ⅳ．平均値の差の検定
Ⅴ．等分散の検定
Ⅵ．比率の検定
Ⅶ．適合度の検定
Ⅷ．独立性の検定

ではあらためて
第13章 Ⅲ．区間推定 の復習から

［復習］ Ⅳ．母平均の区間推定

<u>母平均の区間推定に関するその他の補足</u>
「μの（100×γ）%信頼区間」:

$$\left(\bar{X} - z(\alpha/2)\ \sigma/\sqrt{n}\ ,\ \bar{X} + z(\alpha/2)\ \sigma/\sqrt{n} \right)$$

について:
- 標本平均（の実測値）を中心とする区間である．
- $\sigma/\sqrt{n}$ は標準誤差と呼ばれる．

- γを大きくすると・・・
  - $\alpha = 1-\gamma$ は小さくなる → $z(\alpha/2)$ は大きくなる．
    →信頼区間の幅が大きくなる．
    （外れる確率を減らすのだから、幅を大きく取る必要があるのは当然）
- nを大きくすると → 信頼区間の幅が小さくなる．
  - 情報量が増えるのだから，誤差が減るのは当然

[復習] Ⅳ．母平均の区間推定

「母分散既知の場合」のポイントをまとめると

- 母分散σ^2を用いて標本平均の分布が表せる．
- 母集団分布が正規分布なら標本平均の分布も正規分布．
- 正規母集団を仮定せずとも，標本サイズnが十分大きければ標本平均の分布は正規分布で近似でき，いずれにしても正規分布の問題に帰着できる．
- だがその（標本平均が従う）正規分布$N(\mu, \sigma^2/n)$は，母分散σ^2を用いて表されているのだから，σ^2の値がわからないと推測できない．
- 現実にはσ^2は未知のケースが多い！

一方・・・

Ⅳ．母平均の区間推定

標本の不偏分散 $U^2 := \{ \Sigma (x_i - \bar{x})^2 \} / (n-1)$
は，常にわかる（標本データから計算できる）．

そこで母分散未知の場合には：

- 方針1：U^2をσ^2の近似値として利用．
  - U^2はσ^2の（不偏）推定量ですからね．
  - 「近似」を認めてしまえば，結局またしても正規分布の問題に帰着．推定の方法はほとんど同じ．
    （Uの値をσのところに代入するだけ．同じ公式が使える）
  - 標本サイズが大きければ，良い（誤差の少ない）近似値であることが期待できる．
  - だが小標本の場合は誤差が無視できないはず．
    →方針2へ

Ⅳ．母平均の区間推定（母分散未知）

（ここからが今日の授業の本題）

- 方針2：σ^2を用いずに表せる（U^2を含む）統計量を導入する．

→第11章Ⅴ

3-その他の重要な標本分布 [2]　t分布

（★p.99のt分布の説明冒頭に「XとYが互いに独立な確率変数で、」と追加）

以下の定理を利用．

<u>定理（教科書p.100）</u>

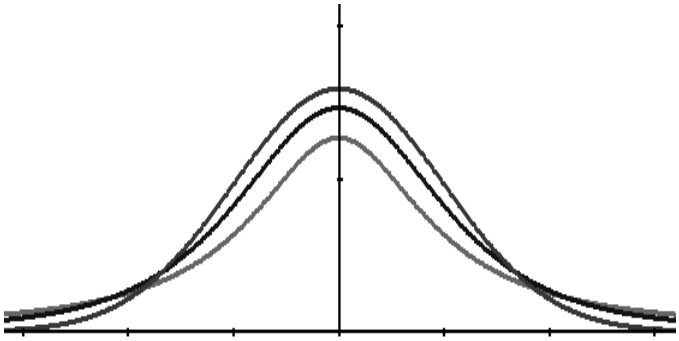
正規母集団$N(\mu, \sigma^2)$からの，大きさnの無作為標本 $\{X_1, X_2, \cdots, X_n\}$について，標本平均X'，不偏分散を$U^2$とする．このとき，確率変数

$$T := (X' - \mu) / (U/\sqrt{n})$$

は自由度n-1のt分布（と呼ばれる分布）に従う．

t分布の密度関数のグラフ

（明星大学 船津好明先生のウェブサイトより転載）



青：自由度3，暗緑：自由度1（赤：標準正規分布）

Ⅳ．母平均の区間推定（母分散未知）

t分布について

- 教科書p.99（第1刷）の定義に「XとYが独立」という条件を追加．
- 「スチューデントのt分布」とも呼ばれる．
  - 「スチューデントstudent」は発案者のペンネーム．
- 厳密に定義するのは少々やっかいなので省略．ともかく理論的に「わかっている」分布として（巻末の数表などを用いて）利用する．
- 密度関数の概形は正規分布と似ている．対称性のある釣鐘型．
- 自由度と呼ばれるパラメータを持つ．
  - サイズnの標本から求めたスチューデント比Tにおいては，自由度ν（ギリシャ文字の「ニュー」）＝n－1．
  - 自由度とは大雑把に言えば「独立なパラメータの個数」．あまりピンとこなくても,当面は「そういうもの」と割り切って利用すればよいです．
  - ν大きいほど標準正規分布N（0,1）に近づき，ν＝∞で完全に一致．
  - 自由度νのt分布の上側100α％点を$t_{\nu}(\alpha)$または$t(\nu, \alpha)$で表す．
  - $t_{\nu}(\alpha)$の値は教科書p.137の表から．

Ⅳ．母平均の区間推定（母分散未知）

注意と補足

- $T=(X'-\mu)/(U/\sqrt{n})$ はスチューデント比と呼ばれる．
- 「正規母集団」の仮定はt分布を利用する上で本質的に重要．
- 標本平均X'の標準化変数：

  $Z=(X'-\mu)/(\sigma/\sqrt{n})$ との類似に注意．
  - つまりnが大きければ，T　Zであるということ．
    （ただし，今考えているのはnが小さいケースでした）

Ⅳ．母平均の区間推定（母分散未知）

t分布を利用したμの推定

- **母分散σ^2が未知の正規母集団$N(\mu, \sigma^2)$からの，大きさnの無作為標本$\{x_1, x_2, \cdots, x_n\}$について，標本平均x'，不偏分散を$U^2$とする．**

このとき

$$t = (x' - \mu) / (U/\sqrt{n}) \sim (\text{自由度}n-1\text{のt分布})$$

だから

- **母平均μの信頼度γ＝1－αの信頼区間は**

$$\left(x' - t_{n-1}(\alpha/2)\, U/\sqrt{n},\ x' + t_{n-1}(\alpha/2)\, U/\sqrt{n} \right)$$

【注意】やってることは母分散既知の場合と大差ない．自由度という新たなパラメータが出てきただけ．

Ⅳ．母平均の区間推定（母分散未知）

例題2（教科書p.108）

- **母集団：成人女性HDLコレステロール値**
  （正規母集団はp.106で仮定）
- **標本サイズ**n＝10，**標本平均**x'＝70
- **標本の不偏分散**＝14×14
  （問題文中の「標準偏差」はここでは不偏分散の平方根）
- **信頼度**＝0.95

［解答］

自由度＝n－1＝10－1＝9

教科書p.137**の表より** $t_9(0.025)$ ＝2.262**を読み取る．**

95%**信頼区間は**

$$\left(x' - t_{n-1}(0.025)\, s/\sqrt{n},\ x' + t_{n-1}(0.025)\, s/\sqrt{n} \right)$$

に該当する値を代入して

$$(70 - 2.26 \times 14/\sqrt{10},\ 70 + 2.26 \times 14/\sqrt{10})$$

$$= (\quad ,\quad)$$

★教科書（第1刷）は要修正！

Ｖ．母分散の区間推定

都合により割愛します
(時間が余りそうだったら後でやる)

→VI. 母比率の区間推定　へ

VI. 母比率の区間推定

「母集団比率に関する推測」とは

- ベルヌーイ母集団に関する推測.
  - 2項分布の応用.
  - ある程度大きな標本を扱うケースがほとんどなので，たいていは「2項分布の正規近似」を利用する.
- 「世論調査の類」. 支持率調査など，身近に興味深い例が多い.
  - 「統計的な理解を深めるよいチャンス」.

VI. 母比率の区間推定

2項分布の正規近似RECALL
中心極限定理により，nが十分大きいとき，
B(n,p) **は**N(np, np(1-p)) **で近似できる．**

∵ B(n,p) **に従う確率変数は，**B(1,p)（**という同一の分布**）**に従う独立な**n**個の確率変数の和と見なせるから．**

- **従って，標本比率**P＝X/n**の分布も，**
  正規分布 N(p, p(1-p)/n) **で近似できる．**
- **さらに**P**の標準化変数**：

$$Z = (P-p)/\sqrt{(p(1-p)/n)}$$

は近似的に標準正規分布に従う．

- **分布が決まれば，あとはこれまでと同じ考え方で進めればよい．．．**

がしかし！

VI. 母比率の区間推定

（**以下母比率を**p**，標本比率**P＝X/n **の実現値を**P_0**とする**）

$$Z = (P-p)/\sqrt{(p(1-p)/n)}$$

を元に$P(Z \in I) = 1-\alpha$**となる区間**I**を構成すると**

$$\left(P_0 - z(\alpha/2)\sqrt{(p(1-p)/n)},\ P_0 + z(\alpha/2)\sqrt{(p(1-p)/n)}\right)$$

となり未知数p**が残ってしまう．そこで**

推定を行う際は，

Z=(P－p)/$\sqrt{}$(p(1－p)/n) **の分母**（P**の標準偏差**）**の**p**を近似値**（**推定値**）P_0**で置き換えて計算．すなわち信頼度**$\gamma = 1-\alpha$**の信頼区間は以下のようになる**：

$$\left(P_0 - z(\alpha/2)\sqrt{(P_0(1-P_0)/n)},\ P_0 + z(\alpha/2)\sqrt{(P_0(1-P_0)/n)}\right)$$

★ただし，誤差の最大値を見積もりたいときはp=0.5 **を採用．**

- p(1－p) **は**p=0.5**のとき最大値**0.25**を取ることに注意（2次関数のグラフを思い出せ！）．**

教科書p.111訂正

［例題5］

誤「回答者330人のうち，18人が」

正「回答者330人のうち，23人が」

（よって標本比率p’＝ 23/330　　0.07となり，
解答はそのまま）

では14章 へ.

From TV News
The Governor of Tokyo said…

“I don’t believe the
opinion poll, because
they surveyed only
1000 people or so.”

（世論調査の結果について
記者から感想を尋ねられ
たのを受けて）

Is 1000 too small?

1000 is :

1000/127,500,000 = 1/127,000 = 0.0008%
of the total population of Japan, and

1000/12,140,000 = 1/1214 = 0.08%
of the total population of Tokyo.

(**確かに母集団と比較すれば小さいが・・・**)

But!

Calculation of SE

- n: sample size
- p: population rate
- SE: Standard Error

SE = $\sqrt{}$(p(1-p)/n)

=< $\sqrt{}$(0.5(1-0.5)/1000) = 0.0158

If p is around 0.1, then

SE = $\sqrt{}$(0.1(1-0.9)/1000) = 0.00949 1%

Estimation of error (誤差の評価)

SE = $\sqrt{}$(p(1-p)/n) depends only on p and n,
i.e., it does not depend on the size of population!

(母集団の大きさは誤差に影響しない！)

- In this case, the value of the error is around 2-3%.
  (計算してみましょう)
- If n=1000×4=4000, the error is reduced by half.
- In many case, the estimation by a sample with n=1000 is reliable. (←無作為抽出がうまくいっていれば)
- On the other hand, we assume random sampling in the above estimation, which can be suspected.

平成14年度試験問題より

「2002年10月1日付けの朝日新聞記事によれば、最近5年間の日本棋院のプロ公式戦15000局において、黒盤勝率は51.86%であった。」

(2) 日本棋院は現行ルールでは先手が有利であるとしてルールの改正方針を決定した。この判断についてどう思うか。仮説検定の考えを用いて考察せよ。

[参考] 朝日新聞の解説 (ちなみに東京版夕刊の1面):
「厳しい勝負の世界ではわずか2%の違いも無視できない」
↑「そりゃ違うだろ！」と突っ込むところ.

- 標本サイズが大きいから誤差は小さく、わずかな差も無視できない (14章で学ぶ仮説検定の用語を用いれば「有意差がある」) のであって、「勝負の厳しさ」とは無関係.